Dossier n°9 : Exemples de traitement d'une série statistique à deux variables numériques. Etude du nuage de points associé : point moyen, corrélation linéaire, ajustement affine, droite de régression.

Rédigé par Cécile COURTOIS, le 27 août 2003
cecile-courtois@wanadoo.fr

I Situation par rapport aux programmes.

L'étude des séries statistiques commence dans les classes de collège et se poursuit au lycée.

En particulier, on introduit, en Terminale ES, les séries statistiques à deux variables numériques et leurs éléments d'étude.

Dans les nouveaux programmes (2002), le coefficient de corrélation linéaire a disparu.

Je choisis donc de situer ce dossier en Terminale ES, anciens et nouveaux programmes.

II Commentaires généraux.

II.1 A propos du sujet.

Comme je l'ai expliqué précédemment, les élèves sont familiarisés avec l'étude de séries statistiques à une variable (valeur du caractère et effectif) depuis le collège.

Toutefois, dans la vie quotidienne et économique, on est parfois amené à rapprocher et à étudier deux caractères (évolution d'une population au cours du temps par exemple).

En particulier, les médias présentent très souvent de telles données.

Il est donc intéressant de pourvoir établir s'il existe un lien de dépendance (ou encore corrélation) entre ces deux caractères ainsi que de savoir si de tels chiffres nous permettent de faire des prévisions sur ce qu'il va se passer.

Il est notamment important que les élèves, futurs citoyens aient un esprit critique face aux chiffres qu'on leur présente.

L'objectif de ce dossier est donc de présenter les principaux éléments d'étude d'une série statistique à deux variables numériques.

II.2 A propos des exercices.

J'ai donc choisi, pour illustrer ce dossier, de vous présenter trois exercices :

- l'exercice n°1 propose une étude classique de série statistique à deux variables ;
- l'exercice n°2 propose de réaliser plusieurs ajustements affines (je préciserai cette notion plus loin) d'une même série statistique ;
- l'exercice n°3 propose un exemple d'ajustement se ramenant à un ajustement affine.

Rappelons, pour cela, les principaux outils se rapportant aux séries statistiques à deux variables.

Dans la suite de ce paragraphe, on s'intéresse à deux variables numériques sur une population $(xi, y_i)_{1 \leq i \leq n}$. On peut modéliser une telle série sous deux formes : un tableau ou un graphique.

<u>Définition 1 :</u>

Dans un repère orthogonal, l'ensemble des points M_i de coordonnées (x_i, y_i), pour i variant de 1 à n, est appelé <u>nuage de points</u> de cette série à deux variables.

Définition 2 :

Le point moyen de ce nuage est le point G de coordonnées ($\bar{x}$; $\bar{y}$) où $\bar{x}$ et $\bar{y}$ sont les moyennes respectives des xi et des yi.

Définition 3 :

On appelle covariance de x et y le nombre noté σ_{xy} (ou C_{xy} ou cov(x,y)) et défini par :

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

On a encore : $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$.

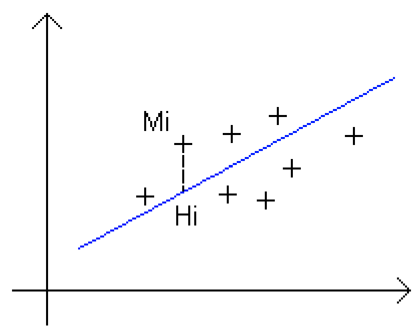
Notion d'ajustement affine :

Parfois, le nuage de points a une forme « allongée » : il semble qu'on peut tracer une droite (même plusieurs) autour de laquelle sont situés les points du nuage. On dit que chacune de ces droites réalise un ajustement affine du nuage.

On peut alors se demander si une de ces droites est « meilleure » que les autres et si oui, selon quel critère.

Théorème 4 :

Il existe une unique droite associée au nuage de points $M_i(x_i ; y_i)$ pour i variant de 1 à n telle que la somme S des $M_iH_i^2$ soit minimale.



Cette droite passe par le point moyen G($\bar{x}$; $\bar{y}$) de ce nuage et a pour équation y = ax+b avec a = $\frac{\sigma_{xy}}{S_x{}^2}$ et b = $\bar{y}$ -a $\bar{x}$.

Définition 5 :

La droite ainsi construite est appelée droite de régression de y en x par la méthode des moindres carrés.

La décision d'ajuster un nuage par une droite semble se prendre à la seule vue du nuage mais les statisticiens ont éprouvé le besoin de quantifier cette prise de décision.

Définition 6 :

On appelle coefficient de corrélation linéaire entre x et y le nombre, noté r, égal à $\frac{\sigma_{xy}}{S_x S_y}$.

La corrélation entre x et y est très forte lorsque $|r| \geq \frac{\sqrt{3}}{2}$. On estime alors que le nuage est suffisamment allongé pour mettre en œuvre un ajustement affine.

Chacun des exercices présentés peut utiliser la calculatrice de façon efficace, notamment pour représenter le nuage de points.

III Présentation des exercices.

III.1 Exercice n°1.

But : Etudier la part du temps partiel au sein de la population active entre 1980 et 1997.

Méthode :

- Représenter le nuage de points et le point moyen ;
- Calculer le coefficient de corrélation linéaire ;
- Déterminer la droite de régression de y en x par la méthode des moindres carrés ;
- La série étant chronologique, la méthode des moindres carrés va permettre de prévoir une situation à venir (en 2004).

Outils :

- Moyenne d'une série statistique ;
- Ecart type d'une série statistique.

III.2 Exercice n°2.

But : Etudier le trafic aérien intérieur français entre 1985 et 1998.

Objectif : Comparer quatre méthodes d'ajustement affine :

- Ajustement par la droite de Mayer (droite passant par les points moyens de deux sous-séries) ;
- Ajustement par une droite arbitraire passant par le point moyen ;
- Ajustement par la droite joignant M_1 et M_n ;
- Ajustement par la méthode des moindres carrés.

Méthode :

- On détermine une équation réduite de chaque droite.
- On utilisera le tableur de la calculatrice pour comparer la somme des résidus S = $\sum_{i=1}^{n}(y_i - ax_i - b)^2$ où les droites ont pour équation y = ax+b.

Commentaire :

Comme il a été annoncé précédemment, la méthode des moindres carrés donnera les meilleurs résultats mais on remarquera que l'ajustement par la droite arbitraire donne également de bons résultats.

Outils :

Equations de droites.

III.3 Exercice n°3.

But : Etudier l'évolution du poids d'un enfant.

Objectif :

Il se peut que la forme du nuage de points ressemble à une fonction connue : $x \to x^2$, $x \to \sqrt{x}$, $x \to e^x$, $x \to \ln x$, etc.

On se ramène alors à une nuage allongé par ce qui s'apparente à un changement de variables. C'est l'objectif de cet exercice.

Commentaire :

Ce dernier exercice propose de prévoir le poids du sujet à 20 ans et 25 ans. On trouve alors un poids de 182 kg (!), ce qui prouve qu'il faut savoir garder un œil attentif et un regard critique sur ces méthodes et études.

IV Enoncés et références des exercices.

IV.1 Exercice n°1 (n°33 p 254, Transmath TES 2002, modifié).

Le tableau suivant, publié en août 1999 dans une revue économique, donne la part du temps partiel au sein de la population française active (les valeurs pour 2000 et 2004 sont le résultat d'une estimation).

Année x_i	1980	1985	1990	1995	1997	2000	2004
Part du temps partiel en % y_i	8,3	11	12	15,6	16,8	18	20

On étudie la série statistique (x_i,y_i) pour x_i compris entre 1980 et 1997.

1. Représenter dans un repère orthogonal le nuage de points de coordonnées $(x_i\,;y_i)$ pour x_i compris entre 1980 et 1997. On prendra 1cm pour une part de 2% en ordonnée et 2cm pour 5 ans en abscisse. L'origine sera le point (1980 ;0).
2. Déterminer les coordonnées de G, point moyen de la série statistique $(x_i\,;y_i)$. Placer ce point sur le graphique.
3. a) Calculer le coefficient de corrélation linéaire r. Expliquer pourquoi un ajustement affine est justifié.

   b) Déterminer l'équation réduite de la droite de régression Δ de y en x par la méthode des moindres carrés.

   c) Tracer cette droite sur le graphique.
4. Peut-on considérer que les estimations pour 2000 et 2004 faites par la revue ont été réalisées en utilisant l'équation obtenue en 3.b) ?

IV.2 Exercice n°2 (n°27 p 117, Déclic TES 2002, modifié).

Le tableau ci-dessous donne le trafic aérien français en milliards de voyageurs-kilomètres.

Année	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
Rang x_i	1	2	3	4	5	6	7	8	9	10	11	12
Trafic y_i	7,4	8,3	8,9	9,6	11	11,4	11,7	12,2	12,3	12,7	12,7	13,8

Année	1997	1998
Rang x_i	13	14
Trafic y_i	13,8	14,5

1. Représenter le nuage de points $M_i(x_i\,;y_i)$ dans un repère orthogonal. Calculer les coordonnées du point moyen G et le placer.
2. Calculer le coefficient de corrélation linéaire. Expliquer pourquoi un ajustement affine est justifié.
3. Ajustement affine n°1 :
   a) Déterminer les coordonnées du point moyen G_1 des sept premiers points du nuage et du point moyen G_2 des sept derniers.
   b) Déterminer l'équation réduite de la droite (G_1G_2) sous la forme y = ax+b. Vérifier que G est un point de (G_1G_2).
4. Ajustement affine n°2 :
   Déterminer l'équation réduite de la droite D passant par G et de coefficient directeur 0,5. Tracer D.
5. Ajustement affine n°3 :
   Calculer l'accroissement moyen annuel du trafic entre 1985 et 1998. En déduire l'équation réduite de la droite (M_1M_{14}). Le point G appartient-il à cette droite ?
6. Ajustement affine n°4 :
   Déterminer l'équation réduite de la droite de régression Δ de y en x par moindres carrés.
7. Comparaison :

A l'aide d'une calculatrice, calculer la somme des résidus $S = \sum_{i=1}^{n}(y_i - ax_i - b)^2$ pour chacune des droites précédentes d'équation réduite de la forme y = ax+b. Quelle est la droite pour laquelle cette somme est minimale ?

IV.3 Exercice n°3 (n°15 p 29, Transmath TES 1998).

Monsieur Otto Math est un papa heureux. Son fils bénéficie d'une excellente santé. Il a noté son poids (en kg) à chacun de ses anniversaires.

Age x_i (en années)	7	8	9	10	11	12
Poids y_i	22	24	28	34	42	51

Soucieux de l'avenir, Otto souhaiterait avoir une idée de l'évolution du poids de son héritier.

1. Représenter cette série par un nuage de points (1cm pour un an en abscisse et 1cm pour 4kg en ordonnée).
2. Posons $z_i = \sqrt{y_i}$ et compléter le tableau précédent avec les z_i arrondis à 10-2 près.
3. a) Sur un autre graphique, représenter les points de coordonnées (x_i ;z_i). Calculer le coefficient de corrélation linéaire entre x et z. Calculer les coordonnées du point moyen G.
   b) Donner une équation réduite de la droite de régression de z en x par moindres carrés.
4. En utilisant cette droite, calculer quel pourrait être le poids de l'héritier à 20 ans et à 25 ans. Que faut-il penser de tels calculs ? Otto doit-il réellement se faire du souci ?

V Commentaires.

On pourra se reporter aux commentaires de la leçon n°7 traitant du même sujet que ce dossier.

J'ai, en particulier, rajouté dans les exercices faisant appel aux nouveaux programmes, le calcul du coefficient de corrélation linéaire. En effet, depuis 2002, la décision d'ajuster un nuage par la méthode des moindres carrés se prend à la seule vue du nuage de points car l'interprétation du coefficient de corrélation linéaire a été jugée trop « délicate » à faire en Terminale ES.

Le dernier exercice, en particulier la dernière question, permet d'éveiller l'esprit critique des élèves et de leur montrer les limites de l'extrapolation par un ajustement affine.